KU LEUVEN

HOGESCHOOL-UNIVERSITEIT BRUSSEL

FACULTEIT ECONOMIE EN BEDRIJFSWETENSCHAPPEN

Probabilistic feature models with individual differences in feature selection

Michel Meulders (KU Leuven)
Jeroen K. Vermunt (Tilburg University)
Iven Van Mechelen (KU Leuven)

FEB@Campus Brussel Research Paper
September 2014
Nr. 2014/02

FACULTEIT ECONOMIE EN BEDRIJFSWETENSCHAPPEN
CAMPUS BRUSSEL (HUBRUSSEL)
Warmoesberg 26
1000 BRUSSEL, België
tel. + 32 2 210 12 11

Probabilistic feature models with individual differences in feature selection

Michel Meulders

KU Leuven

Jeroen K. Vermunt

Tilburg University

Iven Van Mechelen

KU Leuven

Abstract

Probabilistic feature models (PFMs) can be used to explain judgements of persons about binary object-attribute associations on the basis of latent features. More specifically, PFMs assume that persons classify both objects and attributes in terms of binary latent features and that the observed judgement is derived as a non-compensatory (e.g. disjunctive or conjunctive) mapping of the object- and attribute classifications. In this paper we develop multilevel latent class extensions of the PFM that allow to model heterogeneity in the object-attribute association probabilities across persons by assuming that persons select each of the latent features with a class-specific probability when making object-attribute judgements. In addition, statistical dependencies between object-attribute associations with a common element are modelled by assuming that a person relies on the same object classifications for all attribute judgements with regard to that object (or, alternatively, on the same attribute classifications for all object judgements with regard to that attribute). Compared to existing PFM extensions, the model proposed in this paper has several advantages. First, it allows the user to independently specify the number of features and the number of latent person classes, leading to a more flexible modelling. Second, unlike models with class-specific object- or attribute parameters the models presented in this paper use a small set of parameters to model heterogeneity, leading to more stable parameter estimates and models that are easier to interpret. As an illustration, the models are used to analyze data on hostile behavior and psychiatric diagnosis.

Keywords: multilevel latent class model, latent feature, three-way three-mode data

Probabilistic feature models with individual differences in feature selection

Introduction

The analysis of binary three-way data may be of substantive interest in several domains. For instance, in personality psychology one may study the behaviors of persons in different situations in order to identify person types with stable situation-behavior relations (Mischel & Shoda, 1998; Mischel, Shoda, & Mendoza-Denton, 2002; Vansteelandt & Van Mechelen, 1998). In psychiatric diagnosis one may analyze the symptoms assigned to patients by different psychiatrists in order to study the implicit taxonomy used by psychiatrists (Van Mechelen & De Boeck, 1990). In marketing one may study the competitive structure of products by analyzing product-attribute judgements made by consumers (DeSarbo, Grewal, & Scott, 2008; Torres & Bijmolt, 2009). In social network analysis one may study the social structure of a group by analyzing friendship ties between each pair of members as judged by each of the group members (Kumbasar, Kimball, & Batchelder, 1994; González, Tuerlinckx, & De Boeck, 2009).

Probabilistic feature models (PFMs) have been introduced by Maris, De Boeck, and Van Mechelen (1996) to analyze binary three-way data. Furthermore, the models have been applied to several substantive domains such as marketing research (Candel & Maris, 1997; Meulders, 2013), cross-cultural research (Meulders, De Boeck, Van Mechelen, Gelman, & Maris, 2001), and emotion perception (Meulders, De Boeck, Van Mechelen, & Gelman, 2005). More specifically, with data on persons ($i = 1, \ldots, I$) who indicate which behaviors ($k = 1, \ldots, K$) they would display in each of a set of situations ($j = 1, \ldots, J$), PFMs assume that persons classify both situations and behaviors in terms of F *binary latent features* and that these classifications are combined, according to some prespecified mapping rule, to derive the observed judgements. When modelling situation-behavior judgements, the latent features could for instance represent latent situational encodings that may elicit or suppress a certain behavior. For instance, when being angry at someone of higher status, overt aggressive reactions are likely to be suppressed as they are considered

to be inappropriate. Observed judgements are represented using the variable D_{ijk} which equals 1 if person i indicates that she would display behavior k in situation j, and 0 otherwise. The classification of situations is modelled with independent latent Bernoulli variables $X_{ki}^{jf} \sim \text{Bern}(\sigma_{jf})$ which equal 1 if person i perceives feature f in situation j when judging situation-behavior pair (j, k), and 0 otherwise. Likewise, the classification of behaviors is modelled with independent latent Bernoulli variables $Y_{ji}^{kf} \sim \text{Bern}(\rho_{kf})$ which equal 1 if feature f elicits behavior k when person i judges situation-behavior pair (j, k) and 0 otherwise. Finally, PFMs assume that observed situation-behavior associations D_{ijk} are a non-compensatory (i.e., disjunctive or conjunctive) function of situation- and behavior classifications (X_{ki}^{jf} and Y_{ji}^{kf}, $f = 1, \ldots, F$). For instance, using a disjunctive model it is assumed that a person will display a behavior in a situation if the behavior is implied by at least one of the latent features perceived in the situation, or formally, $X_{ki}^{jf} = Y_{ji}^{kf} = 1$ for at least one latent feature f.

A drawback of the basic PFM is that certain model assumptions may be unrealistic in practice. First, as situation- and behavior parameters are the same for all persons, PFMs imply that all persons have the same probability to display a certain behavior in a certain situation. As this assumption may be unrealistic, latent class extensions of the PFM have been developed to model *heterogeneity* in situation- or behavior parameters (Meulders, Tuerlinckx, & Vanpaemel, 2013). Second, as the PFM assumes that each observation is derived from a set of independent Bernoulli variables, it follows that all observations are statistically independent. This assumption may be unrealistic because situation-behavior associations with a common element are likely to be correlated. To solve this problem, Meulders, De Boeck, and Van Mechelen (2003) developed PFMs with adapted stochastic assumptions. In particular to model *dependencies* between situation-behavior judgements with a common situation, one may assume that persons classify situations only once (i.e., the classification of a situation remains constant within persons). Likewise, to account for dependencies between situation-behavior judgements

with a common behavior one may assume that persons classify behaviors only once (i.e., the classification of a behavior remains constant within persons). Finally, Meulders et al. (2013) described multilevel latent class extensions of the PFM that allow to simultaneously model heterogeneity in the situation- and/or behavior parameters and to model dependencies among situation-behavior judgements with a common element.

As an alternative to including class-specific situation- and/or behavior parameters, one may model heterogeneity of the situation-behavior probabilities across persons by assuming that persons consider only a subset of the latent features when making situation-behavior judgements. Meulders, De Boeck, Kuppens, and Van Mechelen (2002) proposed a latent class extension of the PFM in which person classes consider a specific subset of the latent features. In this model, accounting for a certain latent feature is considered to be deterministic conditional on the latent class membership as it is assumed that each latent feature is either considered or not considered by a person. A drawback of the latter model is that the number of latent classes Q is directly related to the number of latent features involved (i.e., $Q = 2^F$), leading to many small latent classes that are hard to interpret if F becomes large. To account for the latter problem Meulders (2011) proposed an alternative latent class extension of the PFM to model heterogeneity in situation-behavior probabilities by assuming that persons, depending on the latent class they belong to, consider each of the latent features with a certain probability when making situation-behavior judgements. In other words, whether or not persons consider a latent feature is assumed to be a probabilistic rather than a deterministic process.

In this paper we present a multilevel extension of the model proposed by Meulders (2011) which makes it possible to account for dependencies between situation-behavior pairs with a common element. The multilevel latent class model presented in this paper has several advantages compared to existing models. First, compared to a model with deterministic feature selection, it allows a more flexible modelling as the number of latent classes and the number of latent features can be independently specified. Second,

compared to a model with class-specific sets of situation- and/or behavior parameters, in a model with class-specific feature weights the number of model parameters increases much less rapidly with the number of latent classes, resulting in models with a better complexity-fit balance (i.e., lower BIC) and that are easier to interpret.

In the following sections, we first present the latent class extension of the PFM in which persons have class-specific probabilities to consider each of the latent features. Second, we develop multilevel extensions of this LC-based PFM to model dependencies between situation-behavior pairs with a common element. Third, we discuss the estimation of the model parameters. Fourth, we will illustrate the models with applications to hostile behavior and psychiatric diagnosis.

Latent class PFM with probabilistic feature selection

To model heterogeneity in the situation-behavior probabilities across persons we propose a model in which persons consider each of the latent features with a certain probability when making situation-behavior judgements. More specifically, the model makes the following assumptions:

1. The latent variable $X_{ik}^{jf} \sim \text{Bern}(\sigma_{jf})$ $(f = 1, \ldots, F)$ equals 1 if latent feature f (i.e., a latent encoding of the situation) is perceived in situation j when person i judges whether she would display behavior k in situation j, and 0 otherwise.

2. The latent variable $Y_{ji}^{kf} \sim \text{Bern}(\rho_{kf})$ $(f = 1, \ldots, F)$ equals 1 if behavior k is elicited by the perception of feature f when person i judges whether she would display behavior k in situation j, and 0 otherwise.

3. It is assumed that, depending on the latent class they belong to, persons have a certain probability to consider a specific latent feature when making judgements. The latent variable G_{iq} $(q = 1, \ldots, Q)$ equals 1 if person i belongs to class q and 0 otherwise. Furthermore the latent variable Z_{ijk}^{f} equals 1 if person i considers feature

f when judging whether she would display behavior k in situation j. It is assumed that $p(z^f_{ijk}|G_{iq}=1) \sim \text{Bern}(\gamma_{qf})$ and that $P(G_{iq}=1)=\xi_q$ with $\sum_q \xi_q = 1$.

4. It is assumed that the observed judgement of a person is obtained as a deterministic mapping of the latent situation and behavior variables and of the specific subset of features considered by the person, that is $D_{ijk} = C(X^{j1}_{ik}, \ldots, X^{jF}_{ik}, Y^{k1}_{ji}, \ldots, Y^{kF}_{ji}, Z^1_{ijk}, \ldots, Z^F_{ijk})$. For instance, using a disjunctive rule, one assumes

$$D_{ijk} = 1 \iff \exists f : X^{jf}_{ki} = Y^{kf}_{ji} = Z^f_{ijk} = 1$$

Assuming a disjunctive mapping rule, the conditional probability that person i of class q will display behavior k in situation j can be derived as follows:

$$\pi_{jkq} = P(D_{ijk}=1|\boldsymbol{\sigma},\boldsymbol{\rho},\boldsymbol{\gamma},G_{iq}=1) \tag{1}$$

$$= \sum_{\mathbf{x}}\sum_{\mathbf{y}}\sum_{\mathbf{z}} p(D_{ijk}=1|\mathbf{x},\mathbf{y},\mathbf{z})p(\mathbf{x}|\boldsymbol{\sigma})p(\mathbf{y}|\boldsymbol{\rho})p(\mathbf{z}|G_{iq}=1,\boldsymbol{\gamma}) \tag{2}$$

$$= 1 - \prod_f (1-\sigma_{jf}\rho_{kf}\gamma_{qf}) \tag{3}$$

Assuming independent persons and independent judgements given latent class membership, the likelihood of the model reads as follows:

$$p(\mathbf{d}|\boldsymbol{\sigma},\boldsymbol{\rho},\boldsymbol{\gamma},\boldsymbol{\xi}) = \prod_i \sum_q \xi_q \prod_j \prod_k (\pi_{jkq})^{d_{ijk}}(1-\pi_{jkq})^{1-d_{ijk}} \tag{4}$$

In what follows, the model defined by (3) and (4) will be denoted as M_1. Note that the basic PFM of Maris et al. (1996) is equivalent to M_1 with $Q=1$ and $\gamma_{1f}=1$ $(f=1,\ldots,F)$.

Modelling statistical dependencies between pairs with a common element

The previously presented model, M_1, is based on the assumption that persons renew situation classifications (i.e., the perception of latent features in the situation) and behavior classifications (i.e., whether or not perceived features elicit a certain behavior) at

each new judgement. These stochastic assumptions imply that, conditional on the latent class membership, all the judgements made by a person are conditionally independent. However, this assumption may be unrealistic as situation-behavior judgements with a common element (pairs with a common behavior or situation) often show stronger dependencies than other situation-behavior pairs (Meulders et al., 2003, 2013).

To model statistical dependencies between situation-behavior pairs with a common situation, one may assume that persons do not renew classifications, but rather use a constant classification of each situation across all related situation-behavior judgements. In other words, one may assume that the observed judgements D_{ijk} $(k = 1, \ldots, K)$ are all based on the same situation classification $X_i^{jf} \sim \text{Bern}(\sigma_{jf})$ $(f = 1 \ldots, F)$. In the same way, statistical dependencies between situation-behavior pairs with a common behavior can be modelled by assuming that persons use a constant classification of each behavior across all related situation-behavior judgements. In particular, one may assume that the observed judgements D_{ijk} $(j = 1, \ldots, J)$ are all based on the same behavior classification $Y_i^{kf} \sim \text{Bern}(\rho_{kf})$ $(f = 1 \ldots, F)$.

For instance, for the probabilistic feature selection model with a *constant* situation classification per person and a *varying* behavior classification, the likelihood reads as follows:

$$p(\mathbf{d}|\boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma}, \boldsymbol{\xi}) = \prod_i \sum_q \xi_q \prod_j \sum_{\mathbf{x}_i^j} \prod_k p(d_{ijk}|\mathbf{x}_i^j, G_{iq} = 1, \boldsymbol{\rho}_k, \boldsymbol{\gamma}_q) p(\mathbf{x}_i^j|\boldsymbol{\sigma}_j) \tag{5}$$

Assuming a disjunctive mapping rule, the conditional probability that person i of class q will display behavior k given the situation classification $\mathbf{x}_i^j$ equals:

$$P(D_{ijk} = 1|\mathbf{x}_i^j, G_{iq} = 1, \boldsymbol{\rho}_k, \boldsymbol{\gamma}_q) = 1 - \prod_f (1 - x_i^{jf} \rho_{kf} \gamma_{qf}). \tag{6}$$

The probabilistic feature selection model with a constant situation classification and a varying behavior classification as defined by (5) and (6) will further be denoted as M_2. Note that a probabilistic feature selection model with a constant classification of behaviors and a varying classification of situations, further denoted as M_3, is similar as it can be

obtained by switching the role of situations and behaviors in M_2.

Finally, we note that M_2 and M_3 can be considered multilevel latent class models (see Vermunt, 2003, 2007) that involve classifications at three levels: First, at highest level, both models involve a classification of persons in Q classes so that persons have a class-specific probability to consider a certain latent feature when making situation-behavior judgements. Second, at the middle level, M_2 involves for each situation a classification of persons in 2^F clusters based on the latent features persons perceive in the situation and M_3 involves for each behavior a classification of persons in 2^F clusters based on the latent features that would elicit the behavior according to the persons. Third, at the lowest level (i.e., the level of the observations D_{ijk}), M_2 involves a classification of behaviors in terms of latent features and M_3 involves a classification of each situation in terms of latent features.

Estimation

For PFMs and their latent class extensions the complete-data likelihood has a simple structure because the observed variables are obtained as a deterministic mapping of Bernoulli distributed latent variables. In particular, the complete-data likelihood reads as follows:

$$p(\mathbf{d}, \mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{g}|\boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma}, \boldsymbol{\xi}) = p(\mathbf{d}|\mathbf{x}, \mathbf{y}, \mathbf{z})p(\mathbf{x}|\boldsymbol{\sigma})p(\mathbf{y}|\boldsymbol{\rho})p(\mathbf{z}|\mathbf{g}, \boldsymbol{\gamma})p(\mathbf{g}|\boldsymbol{\xi})$$

As a result, maximization of the observed incomplete-data (log)likelihood is enhanced by using an EM-algorithm (Dempster, Laird, & Rubin, 1977; Tanner, 1996). Furthermore, in order to guarantee the existence of parameter estimates in the interior of the parameter space (i.e., to prevent boundary estimates), it is convenient to impose a concave prior distribution (Maris et al., 1996; Vermunt & Magidson, 2005, p. 44). More specifically, we

will use the following conjugate prior distribution:

$$\begin{aligned} p(\boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma}, \boldsymbol{\xi}) = & \prod_j \prod_f \text{Beta}(\sigma_{jf} | 1 + \frac{\alpha_\sigma}{J}, 1 + \frac{\beta_\sigma}{J}) \\ & \times \prod_k \prod_f \text{Beta}(\rho_{kf} | 1 + \frac{\alpha_\rho}{K}, 1 + \frac{\beta_\rho}{K}) \\ & \times \prod_q \prod_f \text{Beta}(\gamma_{qf} | 1 + \frac{\alpha_\gamma}{Q}, 1 + \frac{\beta_\gamma}{Q}) \\ & \times \text{Dir}(\boldsymbol{\xi} | 1 + \frac{\delta_1}{Q}, \ldots, 1 + \frac{\delta_Q}{Q}) \end{aligned} \tag{7}$$

Using positive values for the constants α_σ, β_σ, α_ρ, β_ρ, α_γ, β_γ and δ_q $(q = 1 \ldots, Q)$ we obtain a concave prior distribution. In particular, we set $\alpha_\sigma = \beta_\sigma = \alpha_\rho = \beta_\rho = \alpha_\gamma = \beta_\gamma = 1$ and $\delta_q = 2$ $(q = 1 \ldots, Q)$. The complete-data posterior is now defined as:

$$p(\boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma}, \boldsymbol{\xi} | \mathbf{d}, \mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{g}) \propto p(\mathbf{x} | \boldsymbol{\sigma}) p(\mathbf{y} | \boldsymbol{\rho}) p(\mathbf{z} | \boldsymbol{\gamma}, \mathbf{g}) p(\mathbf{g} | \boldsymbol{\xi}) p(\boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma}, \boldsymbol{\xi})$$

As models M_1, M_2 and M_3 are actually constrained (multilevel) latent class models they can be estimated using the syntax module of the standard latent class software Latent GOLD (version 4.5) (Vermunt & Magidson, 2008). In Appendix A, we describe the estimation of the models with Latent GOLD in more detail. In addition, we describe in appendix B a derivation of the EM-algorithm for model M_2. Derivations for the other models are similar.

In addition to locating the posterior mode(s) of the model, it may also be interesting to use a data-augmented Gibbs sampling algorithm to simulate a sample of the observed posterior distribution (Gelfand & Smith, 1990; Tanner & Wong, 1987). In particular, the posterior sample is a rich source of information which supports not only the computation of point estimates of the parameters (e.g. posterior mean), but also the computation of $100(1 - \alpha)\%$ posterior intervals of (any function of) the parameters. These posterior intervals are also valid in small samples whereas standard errors computed in the context of the EM-algorithm are based on asymptotic theory. In Appendix C, we describe a data-augmented Gibbs sampling algorithm for obtaining a sample of the posterior distribution.

Example 1: Analysis of individual differences in hostile behavior

Data

As a first illustration we analyze data of a study on hostile behavior conducted by Vansteelandt and Van Mechelen (1999). In this study 316 persons indicated for all pairs of 4 behaviors and 14 situations to which extent they would display a certain behavior in a certain situation (0=not, 1=limited, 2=strong). In this paper we analyze a subset of 6 situations and 4 behaviors (see also Meulders et al., 2003). Table 1 provides a description of the situations and the behaviors which were taken from on S-R inventory (see also, Endler & Hunt, 1968). To apply probabilistic feature models, the data were dichotomized (0 versus 1 or 2).

Insert Table 1 about here

Analysis

Models M_1, M_2, and M_3 with one up to five latent features ($f = 1 \ldots, 5$) and with one up to five latent classes ($q = 1 \ldots, 5$) were estimated using an EM-algorithm. For each model 20 runs using random starting points were conducted and the solution with the highest posterior density was selected. Table 2 presents fit measures for the five models (out of 75) with the lowest BIC value. Furthermore, for each of the models in Table 2 we use a data-augmented Gibbs sampling algorithm to simulate a sample of the posterior distribution.

Insert Table 2 about here

As can be seen in Table 2, the five models with lowest BIC value all assume a constant behavior classification. To further investigate to which extent the models in Table

2 can capture person differences we compare observed correlations between situation-behavior (SB) pairs with expected correlations under the model. More specifically, we use a posterior predictive check procedure based on 2000 replicated datasets to compute the proportion of correlations among all observed SB-pairs that lie outside their 99% posterior interval (PI) (i.e., p_{all}). Furthermore, as SB-pairs with a common situation or behavior may be expected to show stronger dependencies, we also compute the proportion of correlations outside the 99% PI for SB-pairs with a common situation (i.e., p_{sit}) and with a common behavior (p_{behav}). As indicated by the results in Table 2, all the models can capture dependencies between SB-pairs with a common behavior rather well: only 2% of the correlations lie outside the 99% PI. Note that this result is in line with our expectations as models assuming a fixed behavior classification directly focus on modelling dependencies between SB-pairs with a common behavior. On the other hand, we see in Table 2 that correlations between SB-pairs with a common situation are not fully captured by the models and that increasing the number of latent classes (Q) can help to capture these dependencies somewhat better. More specifically, four-feature models assuming 2, 3 and 4 classes can capture 69%, 81% and 89% of such correlations, respectively. As the model with the lowest BIC can capture correlations among all SB-pairs rather well (94% of the correlations are in their 99% PI), we will discuss the results of this model more in detail. Note that for the selected model, the three classes contain 19%, 25% and 56% of the persons respectively.

Interpretation of the selected model

Figure 1 shows the posterior mean and 95% PI for the parameters of the selected model. In the present application, the features may be interpreted as latent situational encodings that may elicit or suppress a certain behavior (see also Meulders et al., 2003). More specifically, as Feature 1 is likely to be attributed to most situations (except the situation in which you bang your shins against a park bench) it can be interpreted as a

general feature of frustration. Furthermore, this feature elicits the feeling of irritation which is common in all frustrating situations.

Insert Figure 1 about here

Feature 2 reflects the fact that displaying aggressive reactions in a situation is inappropriate because of the presence of another person. Indeed, this feature has a very high probability of being perceived in the situation 'you are unfairly accused of cheating on an examination' where showing aggression would be inappropriate due to the presence of a high status person (i.e., a professor). On the other hand, Feature 2 has a very low probability of being perceived in the situation 'you bang your shins against a park bench' where a verbal aggressive reaction such as cursing is allowed if you are alone. Otherwise, Feature 2 elicits mainly covert reactions to being frustrated such as 'becoming tense' and 'feeling irritated' because overt aggressive reactions would be inappropriate in the situation.

Feature 3 can be interpreted as the fact that someone intentionally wants to hurt your feelings ('you have found out that someone has told lies about you'). Besides overt reactions ('become tense', 'feel irritated'), this feature is likely to elicit both 'cursing' and 'wanting to strike'.

Feature 4 is mainly attributed to situations in which a verbal aggressive reaction is allowed because there is nobody else present in the situation ('you are driving to a party and your car suddenly has a flat tire', 'you are waiting at the bus stop and the bus fails to stop for you', 'you accidentally bang your shins against a park bench'). This feature especially elicits 'cursing' but also 'feelings of irritation'.

As can be seen in the lower part of Figure 1, persons in different classes have different probabilities to attribute a certain feature to a situation, and to display the behaviors which are typically elicited by the feature. In particular, persons in Class 2 are likely to attribute each of the features to a situation. Persons in Class 3 have high probabilities to

attribute each of the features to a situation, except Feature 3. Finally, persons in Class 1 are only likely to attribute Feature 4 to a situation, they have a moderate probability to consider features 1 and 2 and they have a low probability to consider Feature 3.

To further illustrate how different feature selection probabilities per class (γ_{qf}) affect the situation-behavior profiles of these classes, Figures 2, 3, 4 and 5 visualize, for each behavior, the probability that persons in a certain class will display a behavior in a situation. As can be seen in Figures 4 and 5, persons in Class 2 are more likely to curse or strike in any of the investigated situations than persons of other classes. Furthermore, as shown in Figures 2 and 3, in all situations (except 'when you accidentally bang your shins against the park bench') persons in Class 1 are less likely to become tense or to feel irritated than person of other classes.

Insert Figures 2 to 5 about here

Fit of related models

To evaluate the performance of the selected model in terms of model fit, we compare with previously developed PFM extensions. First, we estimate latent class PFMs with one up to five features in which person classes consider a specific subset of latent features (i.e., a deterministic rather than probabilistic feature selection process) (see Meulders et al., 2002). This analysis shows that a deterministic feature selection model with four features has the lowest BIC value (7365), and hence that it is outperformed in terms of BIC by the selected probabilistic feature selection model. Second, we estimate models with one up to five latent features ($f = 1, \ldots, 5$) and one up to five latent classes ($q = 1, \ldots, 5$) with either class-specific situation parameters (σ_{jfq} and ρ_{kf}), or class-specific behavior parameters (σ_{jf} and ρ_{kfq}) and using the same stochastic assumptions as in the selected model (i.e., varying situation classification and constant behavior classification) (see Meulders et al., 2013). The result of this analysis shows that, among models with class-specific situation

parameters, models $(F = 4, T = 1)$ and $(F = 4, T = 2)$ have lowest BIC (7225) and that, among models with class-specific behavior parameters, model $(F = 4, T = 2)$ has lowest BIC (7220). In other words, models with heterogeneity in situation- or behavior parameters fit less well in terms of BIC than the proposed probabilistic feature selection model.

Example 2: Analysis of the structure of psychiatric syndromes

Data

As a second illustration, we analyze data of a study on psychiatric diagnosis gathered by Van Mechelen and De Boeck (1990) (see also Gelman, Van Mechelen, Verbeke, Heitjan, & Meulders, 2005; Leenen, Van Mechelen, De Boeck, & Rosenberg, 1999; Maris et al., 1996). The data consist of the binary judgements of 15 clinicians (psychiatrists and clinical psychologists) who indicated for each of 30 patients and 23 symptoms whether or not a certain patient has a certain symptom. In addition, the clinicians also judged whether or not patients have a substance use-, schizofrenic-, affective- or anxiety disorder. Note that clinicians could attribute more than one disorder to a patient.

Analysis

Maris et al. (1996) used a disjunctive PFM (with stochastic independence assumptions) to explain patient-symptom and patient-disorder associations. In particular, this model assumes that observed judgements are determined by a set of implicit syndromes which are shared by the clinicians. More specifically, when making a patient-symptom judgement, it is assumed that clinicians evaluate, for each of the implicit syndromes, whether the patient suffers from a certain syndrome and whether the symptom is implied by the syndrome. Furthermore, using a disjunctive mapping rule, it is assumed that clinicians assign a symptom to a patient if the patient suffers from at least one of the implicit syndromes that implies the symptom.

Due to the involved stochastic assumptions, the model proposed by Maris et al.

(1996) does not include any individual differences among the clinicians. In this paper we will further investigate whether this assumption is realistic and evaluate whether clinicians may have different probabilities to consider a certain implicit syndrome when making patient-symptom judgements.

Models M_1, M_2 and M_3 with one up to five latent features ($f = 1, \ldots, 5$) and with one up to five latent classes ($q = 1, \ldots, 5$) were estimated using an EM-algorithm. For each model, 20 runs using random starting points were conducted and the solution with the highest posterior density was selected. Table 3 presents fit measures for the five models (out of 75) with lowest BIC. In addition, a one-class five-feature PFM without rater differences is included in the table as a comparison. For each of the models in Table 3, we used a Gibbs sampling algorithm to simulate a sample from the posterior distribution.

As can be seen in Table 3, the three models with lowest BIC are five-feature models that involve varying patient- and varying symptom classification in terms of implicit syndromes and that assume 3, 4 or 5 latent classes with specific feature selection probabilities. A five-feature model without rater differences ($F = 5, Q = 1$) has a considerably higher BIC value than the other models in Table 3 (viz., 10664). Hence the inclusion of class-specific feature selection probabilities clearly improves the global model fit.

As a more specific model check, we evaluate to what extent the models can capture a basic statistic such as the total number of symptoms a clinician assigns to patients, that is, $S_i(\mathbf{d}) = \sum_j \sum_k d_{ijk}$. To evaluate whether this statistic is fitted well by the model we use a posterior predictive check procedure (Gelman, Meng, & Stern, 1996) to simulate the reference distribution of $S_i(\mathbf{d})$ ($i = 1, \ldots, 15$) and we compute the number of clinicians for which the observed value of S_i lies below or above the simulated 99% posterior interval. As can be seen, the models in Table 3 fit the total number of symptoms assigned by clinicians to patients rather well. In particular, model ($F = 5, Q = 3$) underestimates this aspect for two (out of 15) clinicians and the other models in Table 3 underestimate this number for

one clinician only. On the other hand, it turns out that a model without rater differences fails to fit the number of symptoms assigned by clinicians. That is, model $(F = 5, Q = 1)$ underestimates the assigned number of symptoms for 5 (out of 15) clinicians, and it overestimates this aspect for 6 (out of 15) clinicians. As the model with lowest BIC is the most parsimonious one and as it can rather well capture rater differences in the tendency to assign symptoms, we will discuss the results of this model more in detail.

Insert Table 3 about here

Interpretation of the selected model

Figure 6 displays the posterior mean and the 95% posterior interval for the symptom parameters and for the feature selection parameters of the selected model. Note that the last four 'symptoms' in the Figure represent the four existing disorders that were included in the study. As can be seen, some of the implicit syndromes extracted by the model correspond to existing disorders, whereas other implicit syndromes represent a mixture of existing disorders, or isolate symptoms that are not specific to any of the existing disorders. More specifically, Feature 1 matches affective disorder (.98) which is very likely to elicit symptoms of depression (.98), suicide/self mutilation (.69), social isolation (.86) and role impairment (.59). Feature 2 corresponds with substance use disorder (.97), which especially implies symptoms as narcotics/drugs abuse (.93), alcohol abuse (.52) and role impairment (.61). Feature 3 does not correspond to any of the existing disorders that were included in the study. However, this implicit syndrome matches to some extent with 'disruptive, impulse-control and conduct disorders' as it implies symptoms as antisocial (.51), impulse control impairment (.92), and belligerence/negativism (.89). Feature 4 matches schizofrenic disorder, which implies symptoms as speech disorganization (.84), inappropriate affect/behavior (.95), social isolation (.90), disturbance in daily routine (.91), social dullness (.94) and role impairment (.97). Finally, Feature 5 is a mixture of anxiety-

(.83) and affective (.63) disorders . This feature especially implies the symptoms anxiety (.97), role impairment (.87), depression (.64), excessive somatic concerns (.49) and disorientation/memory impairment (.46).

As can be seen in the lower part of Figure 6, clinicians in different classes clearly have different probabilities to consider each of the implicit syndromes when judging patient-symptom associations. More specifically, consideration probabilities for each of the latent syndromes are highest for clinicians of Class 3, intermediate for clinicians of Class 2 and lowest for clinicians of Class 1, indicating that clinicians differ in the number of symptoms they generally tend to assign to patients. Furthermore, clinicians of Class 1 have a very low probability to consider Feature 3 (i.e., implicit syndrome linked to specific symptoms) and Feature 4 (i.e., implicit syndrome related to schizofrenic disorder). As a result, they are less likely to ascribe the symptoms implied by these implicit syndromes to patients.

Insert Figure 6 about here

Discussion

In this paper we presented multilevel latent class extensions of the PFM that allow to model heterogeneity in object-attribute (e.g., situation-behavior) association probabilities by assuming that raters select each of the latent features with a certain probability when making object-attribute judgements. In addition, the model allows to capture dependencies between object-attribute judgements with a common element by assuming a constant object- or attribute classification in terms of latent features. The proposed probabilistic feature selection model has several advantages compared to previously developed (multilevel) latent class PFMs. First, unlike models that assume a deterministic feature

selection process, probabilistic feature selection models allow the user to independently specify the number of rater classes and the number of latent features, leading to a more flexible modelling approach and the selection of models with a better fit-complexity balance (i.e., lower BIC). Second, models with probabilistic feature selection may be an interesting alternative to models that involve class-specific object- and/or attribute parameters because in the former models the number of parameters increases much less rapidly with the number of latent rater classes, leading to the selection of less complex models that are easier to interpret.

Several topics seem worthwhile to consider in future research. First, whereas previous research has mainly focused on the development of latent class extensions of the PFM to model rater heterogeneity in object-attribute association probabilities, it could also be interesting to develop random-effects extensions of the PFM to model heterogeneity in the object-attribute association probability across raters.

Second, in line with previous work on one-feature probabilistic feature models (Meulders, De Boeck, & Van Mechelen, 2001), it could be interesting to further develop confirmatory (multilevel latent class extensions of) probabilistic feature models for applications in which useful design information is available. For instance, imagine situation-behavior data in which seven pairs of behaviors are chosen to respectively measure each of seven behavior types (i.e., anger-out, anger-in, social sharing, avoidance, indirect behavior, assertive behavior, reconciliation) (see Kuppens, Van Mechelen, & Meulders, 2004). In such case, it could be interesting to specify a seven-feature model in which behaviors of a certain type have only a 'loading' on the corresponding latent feature and that the behavior-feature probabilities for the other latent features are constrained to be equal to zero. Another context where confirmatory PFMs could be of interest is in the cognitive assessment of examinees' skills. In particular, Maris (1999) used latent class extensions of conjunctive PFMs to model the responses of examinees to a set of binary items. Such PFMs assume that an examinee can solve an item if she masters each of the

skills required by the item. The model presented by Maris (1999) actually assumes that examinees are classified with respect to the skills they master and that items, in order to be solved, require each of the skills with a certain probability. However, many cognitive diagnostic models used in practice include a skill by item binary incidence Q matrix that specifies for each item which of the skills is required to solve the item (DiBello & Stout, 2007). Confirmatory PFMs could be useful to include the information of the Q matrix in the analysis.

Third, in data from discrete choice experiments it has been recognized that subjects may only attend to specific subsets of attributes when choosing between alternatives, and that failure to account for such attribute processing heterogeneity may lead to an underestimation of marginal willingness-to-pay estimates (Hensher, Rose, & Greene, 2012). The standard approach to model attribute non-attendance is to use a latent class extension of the conditional logit model in which respondents of each latent class only attend to a particular subset of the attributes (Campbell, Hensher, & Scarpa, 2011; Carlsson, Kataria, & Lampi, 2010; Hole, 2011; Collins, Rose, & Hensher, 2013). Similar to latent class PFMs with deterministic feature selection, the standard latent class approach for modelling attribute non-attendance suffers from the fact that the number of latent classes is a direct function of the number of attributes used to define the alternatives, leading to complex models if the number of attributes involved in the experiment increases. Therefore, applying the same idea as in the present paper, an interesting alternative would be to develop a stochastic attribute non-attendance model in which subjects consider each of the attributes with a class-specific probability.

References

Campbell, D., Hensher, D. A., & Scarpa, R. (2011). Non-attendance to attributes in environmental choice analysis: A latent class specification. *Journal of Environmental Planning and Management*, *54*(8), 1061–1076.

Candel, M. J. J. M., & Maris, E. (1997). Perceptual analysis of two-way two-mode frequency data: probability matrix decomposition and two alternatives. *International Journal of Research in Marketing*, *14*, 321–339.

Carlsson, F., Kataria, M., & Lampi, E. (2010). Dealing with ignored attributes in choice experiments on valuation of Sweden's environmental quality objectives. *Environmental and Resource Economics*, *47*, 65–89.

Collins, A. T., Rose, J. M., & Hensher, D. A. (2013). Specification issues in a generalized random parameters attribute nonattendance model. *Transportation Research Part B*, *56*, 234–253.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

DeSarbo, W. S., Grewal, R., & Scott, C. J. (2008). A clusterwise bilinear multidimensional scaling methodology for simultaneous segmentation and positioning analyses. *Journal of Marketing Research*, *45*, 280–292.

DiBello, L. V., & Stout, W. (2007). Guest editors' introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement*, *44*(4), 285–291.

Endler, N. S., & Hunt, J. M. (1968). S-R inventories of hostility and comparisons of the proportions of variance from persons, behaviors, and situations for hostility and anxiousness. *Journal of Personality and Social Psychology*, *9*, 309–315.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *4*, 733–807.

Gelman, A., Van Mechelen, I., Verbeke, G., Heitjan, F., Daniel, & Meulders, M. (2005). Multiple imputation for model checking: Completed-data plots with missing and latent data. *Biometrics*, *61*, 74–85.

González, J., Tuerlinckx, F., & De Boeck, P. (2009). Analyzing structural relations in multivariate dyadic binary data. *Applied Multivariate Research*, *13*, 77–92.

Hensher, D. A., Rose, J. M., & Greene, W. H. (2012). Inferring attribute non-attendance from stated choice data: Implications for willingness to pay estimates and a warning for stated choice experiment design. *Transportation*, *39*, 235–245.

Hole, A. R. (2011). A discrete choice model with endogenous attribute attendance. *Economics Letters*, *110*, 203–205.

Kumbasar, E., Kimball, R. A., & Batchelder, W. H. (1994). Systematic biases in social perception. *American Journal of Sociology*, *100*, 477–505.

Kuppens, P., Van Mechelen, I., & Meulders, M. (2004). Every cloud has a silver lining: Interpersonal and individual differences determinants of anger-related behaviors. *Personality and Social Psychology Bulletin*, *30*, 1550–1564.

Leenen, I., Van Mechelen, I., De Boeck, P., & Rosenberg, S. (1999). INDCLAS: A three-way hierarchical classes model. *Psychometrika*, *64*(1), 9–24.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*, 187–212.

Maris, E., De Boeck, P., & Van Mechelen, I. (1996). Probability matrix decomposition models. *Psychometrika*, *61*, 7–29.

Meulders, M. (2011). Modelling rater differences in the analysis of three-way three-mode binary data. In In W. Gaul, A. Geyer-Schulz, L. Schmidt-Thieme, & J. Kunze (Eds.), *Challenges at the interface of data analysis, computer science, and optimization*, Proceedings of the 34th Annual Conference of the Gesellschaft für Klassifikation e.

V., Karlsruhe, July 21 - 23, 2010. Studies in Classification, Data Analysis, and Knowledge Organization (pp. 223–132). Heidelberg, Berlin: Springer

Meulders, M. (2013). An R package for probabilistic latent feature analysis of two-way two-mode frequencies. *Journal of Statistical Software*, *54*(14), 1–29.

Meulders, M., De Boeck, P., Kuppens, P., & Van Mechelen, I. (2002). Constrained latent class analysis of three-way three-mode data. *Journal of Classification*, *19*, 277–302.

Meulders, M., De Boeck, P., & Van Mechelen, I. (2001). Probability matrix decomposition models and main-effects generalized linear models for the analysis of replicated binary associations. *Computational Statistics & Data Analysis*, *38*, 217–233.

Meulders, M., De Boeck, P., & Van Mechelen, I. (2003). A taxonomy of latent structure assumptions for probability matrix decomposition models. *Psychometrika*, *68*, 61–77.

Meulders, M., De Boeck, P., Van Mechelen, I., & Gelman, A. (2005). Probabilistic feature analysis of facial perception of emotions. *Applied Statistics*, *54*, 781–793.

Meulders, M., De Boeck, P., Van Mechelen, I., Gelman, A., & Maris, E. (2001). Bayesian inference with probability matrix decomposition models. *Journal of Educational and Behavioral Statistics*, *26*, 153–179.

Meulders, M., Tuerlinckx, F., & Vanpaemel, W. (2013). Constrained multilevel latent class models for the analysis of three-way three-mode binary data. *Journal of Classification*, *30*(3), 306–337.

Mischel, W., & Shoda, Y. (1998). Reconciling processing dynamics and personality dispositions. *Annual Review of Psychology*, *49*, 229–258.

Mischel, W., Shoda, Y., & Mendoza-Denton, R. (2002). Situation-behavior profiles as a locus of consistency in personality. *Current Directions in Psychological Science*, *11*, 50–54.

Tanner, M. A. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions* (Third ed.). New York: Springer-Verlag.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528–540.

Torres, A., & Bijmolt, T. H. A. (2009). Assessing brand image through communalities and asymmetries in brand-to-attribute and attribute-to-brand associations. *European Journal of Operational Research*, *195*, 628–640.

Van Mechelen, I., & De Boeck, P. (1990). Projection of binary criterion into a model of hierarchical classes. *Psychometrika*, *55*, 677–694.

Vansteelandt, K. (1999). A formal model for the competency-demand hypothesis. *European Journal of Personality*, *13*, 429–442.

Vansteelandt, K., & Van Mechelen, I. (1998). Individual differences in situation-behavior profiles: A triple typology model. *Journal of Personality and Social Psychology*, *75*, 751–765.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology*, *33*, 213–239.

Vermunt, J. K. (2007). A hierarchical mixture model for clustering three-way data sets. *Computational Statistics & Data Analysis*, *51*, 5368–5376.

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for latent gold 4.0: Basic and advanced*. Belmont Massachusetts: Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2008). *Lg-syntax user's guide: Manual for latent gold 4.5 syntax module*. Belmont Massachusetts: Statistical Innovations Inc.

Appendix A: syntax code for estimation with Latent GOLD

Data

To illustrate the estimation of models with Latent GOLD we consider the application of M_2 on situation-behavior judgements of raters for all pairs of 2 situations and 5 behaviors. To analyze the data with Latent GOLD one should structure the data as in Table 4. The data should be sorted by person and within person by situation because M_3 involves both a classification at the person level and a classification at the person-situation level.

Insert Table 4 about here

Syntax code

1. model
2. title M2-Q3-F2;
3. options
4. algorithm tolerance=1e-010 emtolerance=0.01 emiterations=250 nriterations=50;
5. startvalues seed=0 sets=20 tolerance=1e-008 iterations=50;
6. bayes categorical=1 variances=1 latent=2 poisson=1;
7. output estimatedvalues parameters=last standarderrors probmeans=posterior profile bivariateresiduals;
8. variables
9. groupID person;
10. caseID situation;

11. dependent D nominal;

12. independent situation nominal, behavior nominal;

13. latent G group nominal 3, X1 nominal 2, X2 nominal 2, Y1 nominal 2 dynamic, Y2 nominal 2 dynamic, Z1 nominal 2 dynamic, Z2 nominal 2 dynamic;

14. equations

15. G <– 1;

16. X1 <– situation;

17. Y1 <– behavior;

18. X2 <– situation;

19. Y2 <– behavior;

20. Z1 <– G;

21. Z2 <– G;

22. D <– (b) 1 + (c) 1| X1 Y1 Z1 + (c) 1| X2 Y2 Z2;

23. b = -100;

24. c = 0;

25. c[8] = 200;

26. end model

Comments

Lines 3-10 contain specific options used for estimation by Latent GOLD (for more information see Vermunt & Magidson, 2008). Lines 9 and 10 indicate that the model involves both a classification at the person level and a classification at the person-situation level. Lines 11 and 12 are used to describe dependent and independent observed variables. Line 13 is used to describe the latent variables and their classification level: The latent variable G with the keyword 'group' is used to classify persons (=groupID variable) in 3 classes, the latent variables $X1$ and $X2$ are used to classify persons for each situation (= caseID variable) in 2×2 clusters, and the latent variables $Y1$, $Y2$, $Z1$ and $Z2$ with the keyword 'dynamic' are used to classify persons in $2^4 = 16$ classes at the level of the individual observations D. Lines 14-22 are used to describe the relationships between the variables involved in the analysis. For instance, lines 20-21 indicate that the probability to select a certain latent feature depends on the latent class the person belongs to. Line 22 indicates that observations D are obtained as a mapping of the latent variables $X1$, $X2$, $Y1$, $Y2$, $Z1$ and $Z2$. In particular, $P(D = 1|\mathbf{x}, \mathbf{y}, \mathbf{z})$ is put equal to $\exp(b+c)/(1+\exp(b+c))$ with $b+c$ being a large positive number for latent data patterns with $X1 = Y1 = Z1 = 1$ or $X2 = Y2 = Z2 = 1$ and with $b+c$ being a large negative number for all other latent data patterns (see lines 23-25).

Appendix B: Computation of the posterior mode

In this section we describe as an example the derivation of the EM-algorithm for model M_2. For this model, the complete-data likelihood is proportional to:

$$\prod_i \prod_q (\xi_q)^{g_{iq}}$$
$$\times \prod_i \prod_j \prod_f (\sigma_{jf})^{x_i^{jf}} (1-\sigma_{jf})^{(1-x_i^{jf})}$$
$$\times \prod_i \prod_k \prod_f \prod_j (\rho_{kf})^{y_{ji}^{kf}} (1-\rho_{kf})^{(1-y_{ji}^{kf})}$$
$$\times \prod_i \prod_j \prod_k \prod_q \prod_f (\gamma_{qf})^{z_{ijk}^f g_{iq}} (1-\gamma_{qf})^{(1-z_{ijk}^f) g_{iq}}$$

Using the prior in (7) with $\alpha_\sigma = \beta_\sigma = \alpha_\rho = \beta_\rho = \alpha_\gamma = \beta_\gamma = 1$ and $\delta_q = 2$ $(q = 1 \ldots, Q)$, except for a constant, the logarithm of the complete-data posterior reads as follows:

$$
\begin{aligned}
&= \sum_q (\frac{2}{Q} + \sum_i g_{iq}) \ln(\xi_q) \\
&+ \sum_j \sum_f [\frac{1}{J} + \sum_i x_i^{jf}] \ln(\sigma_{jf}) + [\frac{1}{J} + \sum_i (1 - x_i^{jf})] \ln(1 - \sigma_{jf}) \\
&+ \sum_k \sum_f [\frac{1}{K} + \sum_i \sum_j y_{ji}^{kf}] \ln(\rho_{kf}) + [\frac{1}{K} + \sum_i \sum_j (1 - y_{ji}^{kf})] \ln(1 - \rho_{kf}) \\
&+ \sum_q \sum_f [\frac{1}{Q} + \sum_i \sum_j \sum_k z_{ijk}^f g_{iq}] \ln(\gamma_{qf}) + [\frac{1}{Q} + \sum_i \sum_j \sum_k (1 - z_{ijk}^{qf}) g_{iq}] \ln(1 - \gamma_{qf})
\end{aligned}
$$

Using EM to estimate the parameters $\boldsymbol{\theta} = (\boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma}, \boldsymbol{\xi})$ means that we maximize the (logarithm of) the complete-data posterior as a function of the parameters, and that we replace the complete data statistics by their conditional expected values given the observed data. Maximization of the posterior distribution yields:

$$\hat{\sigma}_{jf} = \frac{\frac{1}{J} + E(\sum_i x_i^{jf} | \mathbf{d}, \boldsymbol{\theta})}{\frac{2}{J} + I} \tag{8}$$

$$\hat{\rho}_{kf} = \frac{\frac{1}{K} + E(\sum_i \sum_j y_{ji}^{kf} | \mathbf{d}, \boldsymbol{\theta})}{\frac{2}{K} + IJ} \tag{9}$$

$$\hat{\gamma}_{qf} = \frac{\frac{1}{Q} + E(\sum_i \sum_j \sum_k z_{ijk}^f g_{iq} | \mathbf{d}, \boldsymbol{\theta})}{\frac{2}{Q} + (JK) E(\sum_i g_{iq} | \mathbf{d}, \boldsymbol{\theta})} \tag{10}$$

$$\hat{\xi}_q = \frac{\frac{2}{Q} + E(\sum_i g_{iq} | \mathbf{d}, \boldsymbol{\theta})}{I + 2} \tag{11}$$

To update ξ_q one has to compute the conditional expected value of $\sum_i g_{iq}$ which equals:

$$E(\sum_i G_{iq} | \mathbf{d}_i, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma}, \boldsymbol{\xi}) = \sum_i P(G_{iq} = 1 | \mathbf{d}_i, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma}, \boldsymbol{\xi}).$$

The posterior probability $P(G_{iq} = 1|\mathbf{d}_i, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma}, \boldsymbol{\xi})$ can be computed as follows:

$$= \frac{P(G_{iq} = 1|\boldsymbol{\xi})p(\mathbf{d}_i|G_{iq} = 1, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma})}{p(\mathbf{d}_i|\boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma}, \boldsymbol{\xi})} \tag{12}$$

$$\propto \xi_q[\prod_j p(\mathbf{d}_{ij}|G_{iq} = 1, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma})] \tag{13}$$

$$\propto \xi_q \prod_j \{\sum_{\mathbf{x}_i^j} p(\mathbf{d}_{ij}|\mathbf{x}_i^j, G_{iq} = 1, \boldsymbol{\rho}, \boldsymbol{\gamma}) p(\mathbf{x}_i^j|\boldsymbol{\sigma}_j)\} \tag{14}$$

$$\propto \xi_q \prod_j \{\sum_{\mathbf{x}_i^j} [\prod_k p(d_{ijk}|\mathbf{x}_i^j, G_{iq} = 1, \boldsymbol{\rho}_k, \boldsymbol{\gamma}_q)] p(\mathbf{x}_i^j|\boldsymbol{\sigma}_j)\} \tag{15}$$

with

$$p(d_{ijk}|\mathbf{x}_i^j, G_{iq} = 1, \boldsymbol{\rho}_k, \boldsymbol{\gamma}_q) = [1 - \prod_f (1 - x_i^{jf} \rho_{kf} \gamma_{qf})]^{d_{ijk}} [\prod_f (1 - x_i^{jf} \rho_{kf} \gamma_{qf})]^{1 - d_{ijk}}$$

and

$$p(\mathbf{x}_i^j|\boldsymbol{\sigma}_j) = \prod_f (\sigma_{jf})^{x_i^{jf}} (1 - \sigma_{jf})^{1 - x_i^{jf}}$$

To update $\hat{\sigma}_{jf}$ we compute the conditional expected value of $\sum_i x_i^{jf}$ as follows:

$$= E(\sum_i x_i^{jf}|\mathbf{d}_i, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma}, \boldsymbol{\xi})$$

$$= \sum_i \sum_q P(G_{iq} = 1, X_i^{jf} = 1|\mathbf{d}_i, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma}, \boldsymbol{\xi})$$

$$= \sum_i \sum_q P(G_{iq} = 1|\mathbf{d}_i, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma}, \boldsymbol{\xi}) P(X_i^{jf} = 1|G_{iq} = 1, \mathbf{d}_{ij}, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma})$$

$$= \sum_i \sum_q P(G_{iq} = 1|\mathbf{d}_i, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma}, \boldsymbol{\xi}) \sum_{\mathbf{x}_i^j} x_i^{jf} P(\mathbf{X}_i^j = \mathbf{x}_i^j|G_{iq} = 1, \mathbf{d}_{ij}, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma})$$

with

$$P(\mathbf{X}_i^j = \mathbf{x}_i^j|G_{iq} = 1, \mathbf{d}_{ij}, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma}) \propto \left[\prod_k p(d_{ijk}|\mathbf{x}_i^j, G_{iq} = 1, \boldsymbol{\rho}_k, \boldsymbol{\gamma}_q)\right] p(\mathbf{x}_i^j|\boldsymbol{\sigma}_j)$$

To update $\hat{\rho}_{kf}$ we compute the conditional expected value of $\sum_i \sum_j y_{ji}^{kf}$ as follows:

$$\sum_i \sum_q P(G_{iq} = 1|\mathbf{d}_i) \sum_j \sum_{\mathbf{x}_i^j} p(\mathbf{x}_i^j|\mathbf{d}_{ij}, G_{iq} = 1) \sum_{\mathbf{z}_{ijk}} \sum_{\mathbf{y}_{ji}^k} y_{ji}^{kf} p(\mathbf{y}_{ji}^k, \mathbf{z}_{ijk}|G_{iq} = 1, \mathbf{x}_i^j, d_{ijk})$$

with

$$p(\mathbf{y}_{ji}^k, \mathbf{z}_{ijk}|G_{iq} = 1, \mathbf{x}_i^j, d_{ijk}) = \frac{p(d_{ijk}|\mathbf{x}_i^j, \mathbf{y}_{ji}^k, \mathbf{z}_{ijk}) p(\mathbf{y}_{ji}^k|\boldsymbol{\rho}) p(\mathbf{z}_{ijk}|G_{iq} = 1, \boldsymbol{\gamma})}{p(d_{ijk}|\mathbf{x}_i^j, G_{iq} = 1)}$$

Finally, to update $\hat{\gamma}_{qf}$ we compute the conditional expected value of $\sum_i \sum_j \sum_k z_{ijk}^f g_{iq}$ in the numerator of (10) as follows:

$$\sum_i P(G_{iq}=1|\mathbf{d}_i) \sum_j \sum_{\mathbf{x}_i^j} p(\mathbf{x}_i^j|\mathbf{d}_{ij}, G_{iq}=1) \sum_k \sum_{\mathbf{y}_{ji}^k} \sum_{\mathbf{z}_{ijk}} z_{ijk}^f p(\mathbf{y}_{ji}^k, \mathbf{z}_{ijk}|G_{iq}=1, \mathbf{x}_i^j, d_{ijk})$$

Appendix C: computation of sample of the posterior distribution

In order to estimate a sample of the observed posterior distribution $p(\boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma}, \boldsymbol{\xi}|\mathbf{d})$ for model M_2 one can use a Gibbs sampling algorithm which iterates between the following steps:

1. For each entity i draw the vector $\mathbf{g}_i$ from

$$p(\mathbf{g}_i|\mathbf{d}_i, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma}, \boldsymbol{\xi}) \propto p(\mathbf{d}_i|\mathbf{g}_i, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma}) p(\mathbf{g}_i|\boldsymbol{\xi})$$

with

$$p(\mathbf{d}_i|\mathbf{g}_i, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma}) = \prod_j \sum_{\mathbf{x}_i^j} [\prod_k p(d_{ijk}|\mathbf{g}_i, \mathbf{x}_i^j, \boldsymbol{\rho}_k, \boldsymbol{\gamma})] p(\mathbf{x}_i^j|\boldsymbol{\sigma}_j) \tag{16}$$

$$p(\mathbf{g}_i|\boldsymbol{\xi}) = \prod_q \xi_q^{g_{iq}} \tag{17}$$

To compute (16) we use

$$p(d_{ijk}|\mathbf{g}_i, \mathbf{x}_i^j, \boldsymbol{\rho}_k, \boldsymbol{\gamma}) = \prod_q \left\{ \left[1 - \prod_f (1 - x_i^{jf} \rho_{kf} \gamma_{qf}) \right]^{d_{ijk}} \left[\prod_f (1 - x_i^{jf} \rho_{kf} \gamma_{qf}) \right]^{1-d_{ijk}} \right\}^{g_{iq}}$$

$$p(\mathbf{x}_i^j|\boldsymbol{\sigma}_j) = \prod_f (\sigma_{jf})^{x_i^{jf}} (1 - \sigma_{jf})^{1-x_i^{jf}}$$

2. For each pair (i, j), draw $\mathbf{x}_i^j$ from

$$p(\mathbf{x}_i^j|\mathbf{d}_{ij}, \mathbf{g}_i, \boldsymbol{\sigma}_j, \boldsymbol{\rho}, \boldsymbol{\gamma}) \propto p(\mathbf{d}_{ij}|\mathbf{x}_i^j, \mathbf{g}_i, \boldsymbol{\rho}, \boldsymbol{\gamma}) p(\mathbf{x}_i^j|\boldsymbol{\sigma}_j)$$

with

$$p(\mathbf{d}_{ij}|\mathbf{x}_i^j, \mathbf{g}_i, \boldsymbol{\rho}, \boldsymbol{\gamma}) = \prod_q \left\{ \prod_k \left[1 - \prod_f (1 - x_i^{jf} \rho_{kf} \gamma_{qf}) \right]^{d_{ijk}} \left[\prod_f (1 - x_i^{jf} \rho_{kf} \gamma_{qf}) \right]^{1-d_{ijk}} \right\}^{g_{iq}}$$

$$p(\mathbf{x}_i^j|\boldsymbol{\sigma}_j) = \prod_f (\sigma_{jf})^{x_i^{jf}} (1 - \sigma_{jf})^{1-x_i^{jf}}$$

3. For each triple (i, j, k), draw $\mathbf{y}_{ji}^k$ from

$$p(\mathbf{y}_{ji}^k|d_{ijk}, \mathbf{g}_i, \mathbf{x}_i^j, \boldsymbol{\rho}_k, \boldsymbol{\gamma}) \propto p(d_{ijk}|\mathbf{x}_i^j, \mathbf{y}_{ji}^k, \mathbf{g}_i, \boldsymbol{\gamma}) p(\mathbf{y}_{ji}^k|\boldsymbol{\rho}_k)$$

with

$$p(d_{ijk}|\mathbf{x}_i^j, \mathbf{y}_{ji}^k, \mathbf{g}_i, \boldsymbol{\gamma}) = \prod_q \left\{ \left[1 - \prod_f (1 - x_i^{jf} y_{ji}^{kf} \gamma_{qf}) \right]^{d_{ijk}} \left[\prod_f (1 - x_i^{jf} y_{ji}^{kf} \gamma_{qf}) \right]^{1-d_{ijk}} \right\}^{g_{iq}}$$

$$p(\mathbf{y}_{ji}^k|\boldsymbol{\rho}_k) = \prod_f (\rho_{kf})^{y_{ji}^{kf}} (1 - \rho_{kf})^{1-y_{ji}^{kf}}$$

4. for each triple (i, j, k), draw $\mathbf{z}_{ijk}$ from

$$p(\mathbf{z}_{ijk}|\mathbf{x}_i^j, \mathbf{y}_{ji}^k, d_{ijk}, \mathbf{g}_i, \boldsymbol{\gamma}) \propto p(d_{ijk}|\mathbf{x}_i^j, \mathbf{y}_{ji}^k, \mathbf{z}_{ijk}) p(\mathbf{z}_{ijk}|\mathbf{g}_i, \boldsymbol{\gamma})$$

with

$$p(d_{ijk}|\mathbf{x}_i^j, \mathbf{y}_{ji}^k, \mathbf{z}_{ijk}) = \prod_q \left\{ \left[1 - \prod_f (1 - x_i^{jf} y_{ji}^{kf} z_{ijk}^f) \right]^{d_{ijk}} \left[\prod_f (1 - x_i^{jf} y_{ji}^{kf} z_{ijk}^f) \right]^{1-d_{ijk}} \right\}^{g_{iq}}$$

$$p(\mathbf{z}_{ijk}|\mathbf{g}_i, \boldsymbol{\gamma}) = \prod_q \prod_f \left[(\gamma_{qf})^{z_{ijk}^f} (1 - \gamma_{qf})^{1-z_{ijk}^f} \right]^{g_{iq}}$$

5. For each pair (j, f) draw σ_{jf} from

$$\text{Beta}\left(\alpha_\sigma + \sum_i x_i^{jf}, \beta_\sigma + \sum_i (1 - x_i^{jf}) \right)$$

6. For each pair (k, f) draw ρ_{kf} from

$$\text{Beta}\left(\alpha_\rho + \sum_i \sum_j y_{ji}^{kf}, \beta_\rho + \sum_i \sum_j (1 - y_{ji}^{kf}) \right)$$

7. For each pair (q, f) draw γ_{qf} from

$$\text{Beta}\left(\alpha_\gamma + \sum_i \sum_j \sum_k z_{ijk}^f g_{iq}, \beta_\gamma + \sum_i \sum_j \sum_k (1 - z_{ijk}^f) g_{iq}\right)$$

8. Draw $\boldsymbol{\xi}$ from

$$\text{Dirichlet}\left(\delta_1 + \sum_i g_{i1}, \ldots, \delta_Q + \sum_i g_{iQ}\right)$$

It can be shown that the subsequent draws $(\boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\gamma}, \boldsymbol{\xi})$ form a Markov chain which converges towards the true posterior distribution (Tanner & Wong, 1987).

Table 1

Description of situations and behaviors

type	element
situation	your instructor unfairly accuses you of cheating on an examination
	you have just found out that someone has told lies about you
	you are driving to a party and suddenly your car has a flat tire
	you are waiting at the bus stop and the bus fails to stop for you
	someone has opened your personal mail
	you accidentally bang your shins against a park bench
behavior	become tense
	feel irritated
	curse
	want to strike something or someone

From "S-R Inventories of Hostility and Comparisons of the Proportions of Variance from Persons, Behaviors, and Situations for Hostility and Anxiousness" by N.S. Endler and J.M. Hunt, 1968, Journal of Personality and Social Psychology, 9, pp. 310-311. Copyright 1968 by the American Psychological Association. Adapted by permission of the author.

Table 2

Number of features (F), number of latent classes (Q), type of situation classification, type of behavior classification, loglikelihood (LL), number of parameters (Npar), BIC value, proportion of observed correlations between situation-behavior pairs outside the 99% PI (p_{all}), proportion of observed correlations between situation-behavior pairs with a common situation outside the 99% PI (p_{sit}) and proportion of observed correlations between situation-behavior pairs with a common behavior outside the 99% PI (p_{behav}), for the five models with lowest BIC.

F	Q	situation classification	behavior classification	LL	Npar	BIC	p_{all}	p_{sit}	p_{behav}
4	3	varying	constant	-3445	54	7200	.06	.19	.02
4	2	varying	constant	-3462	49	7206	.10	.31	.03
4	4	varying	constant	-3434	59	7207	.05	.11	.02
5	4	varying	constant	-3407	73	7217	.04	.14	.02
5	3	varying	constant	-3416	67	7218	.04	.11	.02

Table 3

Number of features (F), number of latent classes (Q), type of patient classification (patient), type of symptom classification (symptom), loglikelihood (LL), number of parameters (Npar), BIC value, number of clinicians for which S_i lies below (N_{below}) or above (N_{above}) the 99% PI for the five models with the lowest BIC (first five models in the table) and for a model without rater differences (last model in the table)

F	Q	patient classification	symptom classification	LL	Npar	BIC	N_{below}	N_{above}
5	3	varying	varying	-4738	302	10295	2/15	0/15
5	5	varying	varying	-4722	314	10295	1/15	0/15
5	4	varying	varying	-4736	308	10306	1/15	0/15
5	5	varying	varying	-4735	314	10320	1/15	0/15
5	4	varying	varying	-4743	308	10320	1/15	0/15
5	1	varying	varying	-4939	290	10664	5/15	6/15

Table 4

Data structure for analysis with Latent GOLD 4.5. Observed judgements (D) are stored in different records. The variables person, situation and behavior are added to describe each observation.

person	situation	behavior	D
1	1	1	1
1	1	2	0
1	1	3	1
1	1	4	1
1	1	5	0
1	2	1	1
1	2	2	0
1	2	3	0
1	2	4	0
1	2	5	1
2	1	1	0
...			

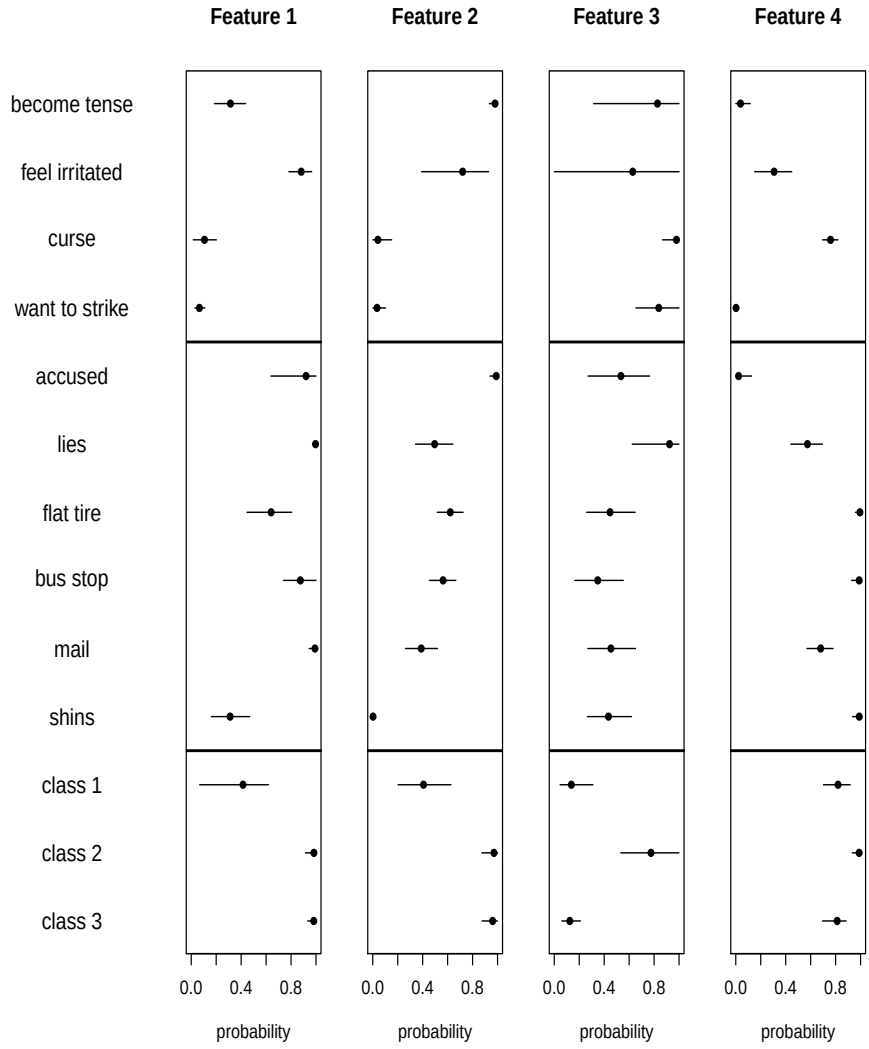

Figure 1. Posterior mean and 95% posterior interval for the situation parameters $\boldsymbol{\sigma}$ (upper part figure), behavior parameters $\boldsymbol{\rho}$ (middle part figure) and selection parameters $\boldsymbol{\gamma}$ (bottom part figure) of a three-class four-feature model ($Q = 3, F = 4$) with probabilistic feature selection, varying situation classification and constant behavior classification.

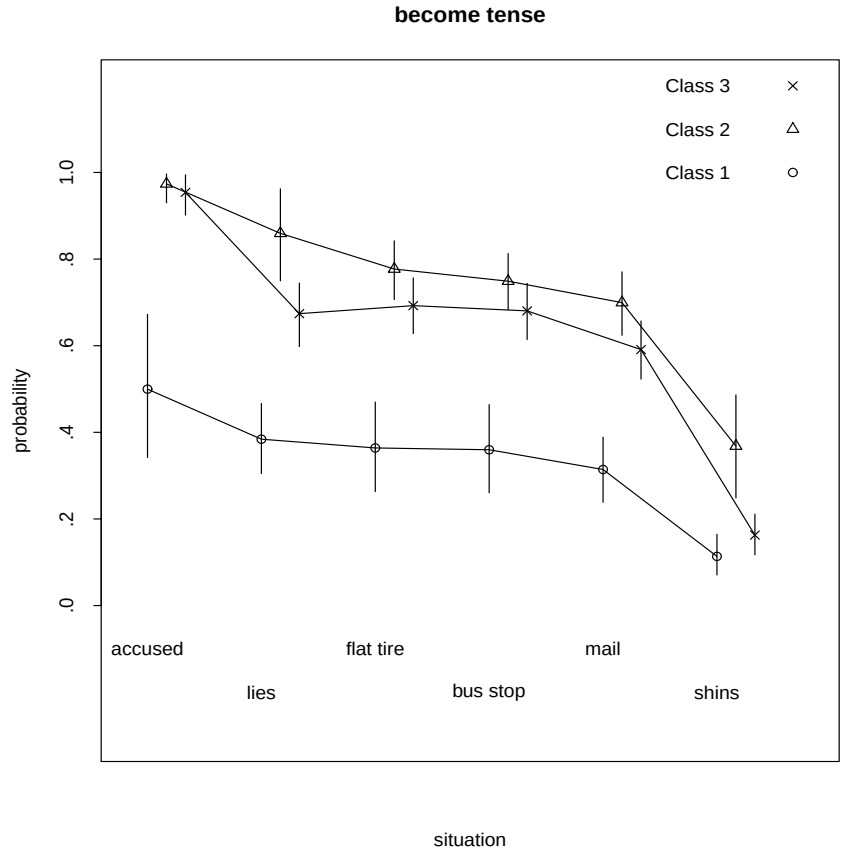

Figure 2. Posterior mean and 95% posterior interval of the class-specific probability to become tense in a situation for the three-class four-feature model ($Q = 3, F = 4$) with probabilistic feature selection, varying situation classification and constant behavior classification.

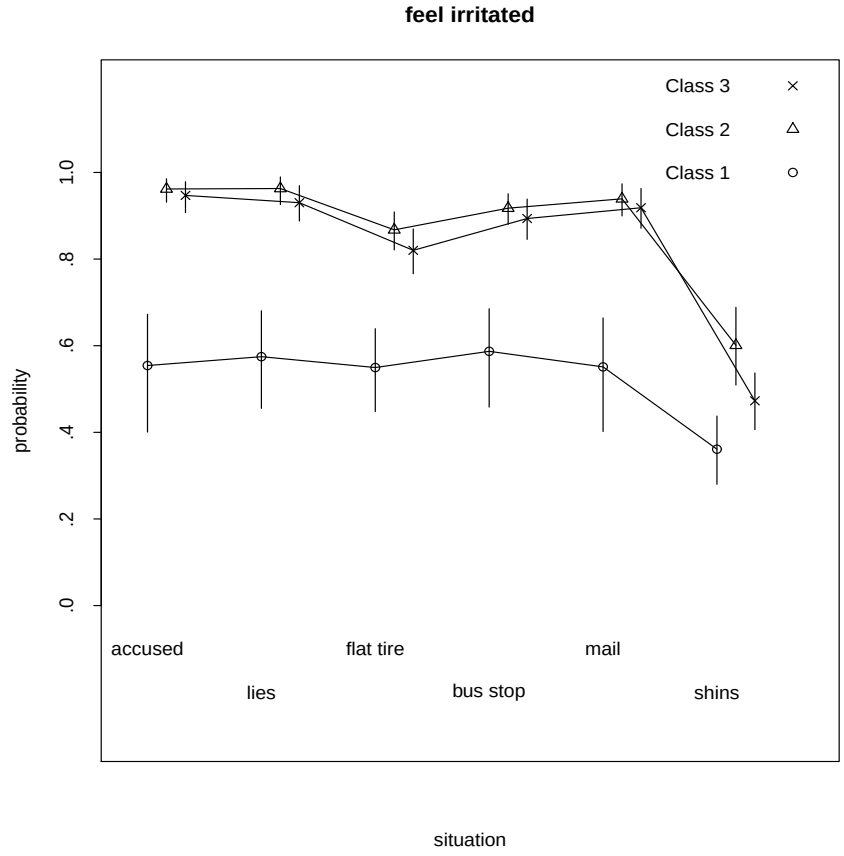

Figure 3. Posterior mean and 95% posterior interval of the class-specific probability to feel irritated in a situation for the three-class four-feature model ($Q = 3, F = 4$) with probabilistic feature selection, varying situation classification and constant behavior classification.

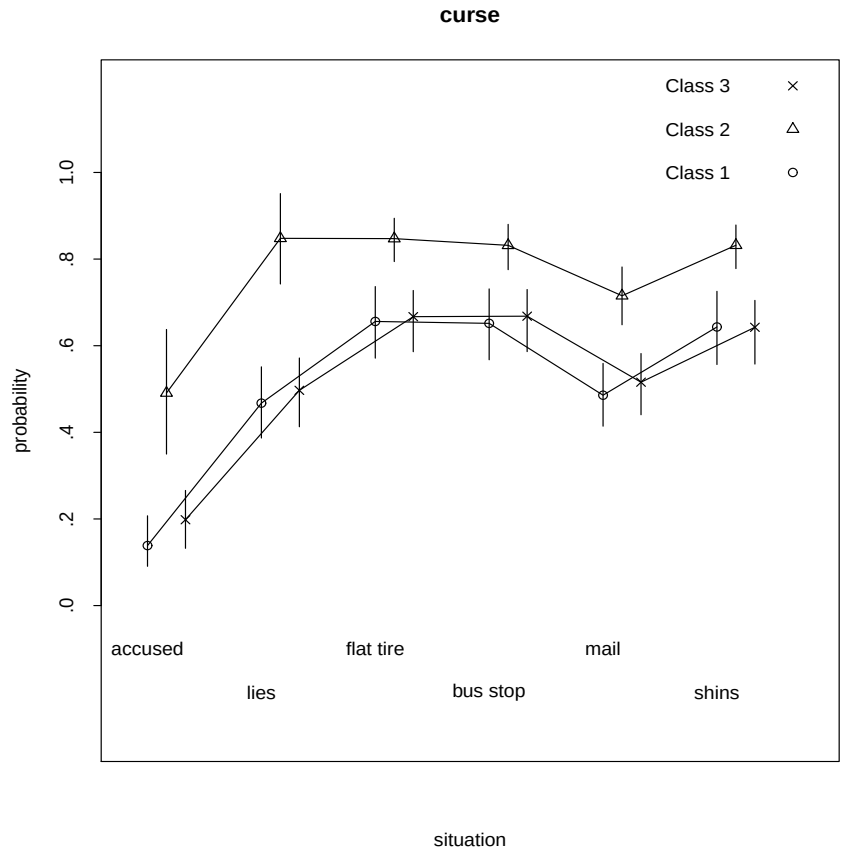

Figure 4. Posterior mean and 95% posterior interval of the class-specific probability to curse in a situation for the three-class four-feature model ($Q = 3, F = 4$) with probabilistic feature selection, varying situation classification and constant behavior classification.

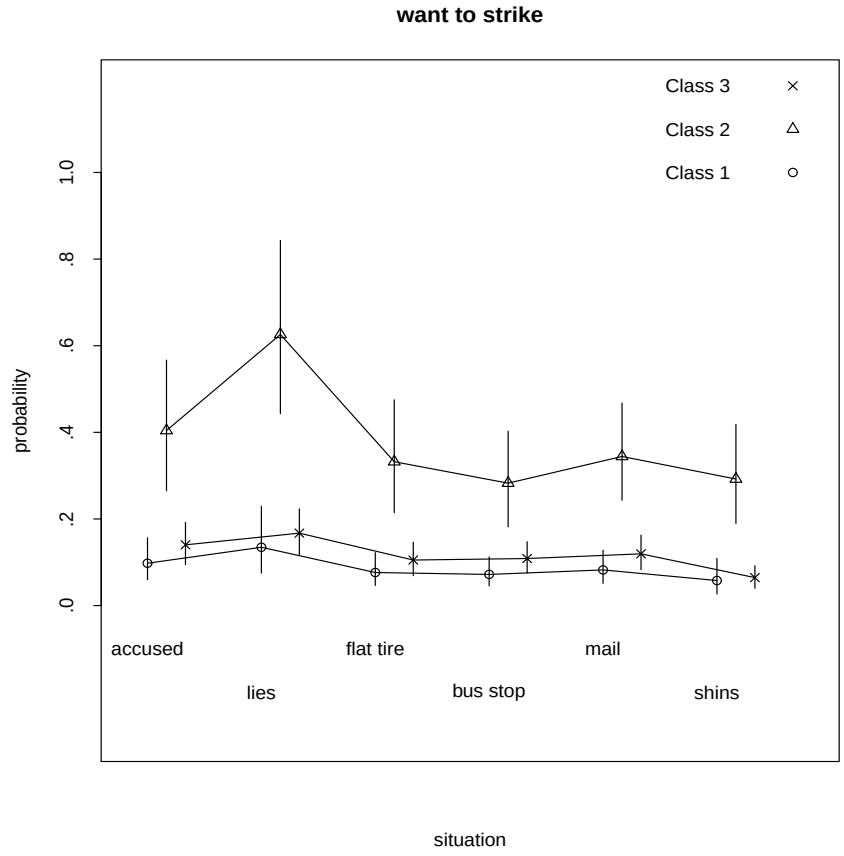

Figure 5. Posterior mean and 95% posterior interval of the class-specific probability of wanting to strike in a situation for the three-class four-feature model ($Q = 3, F = 4$) with probabilistic feature selection, varying situation classification and constant behavior classification.

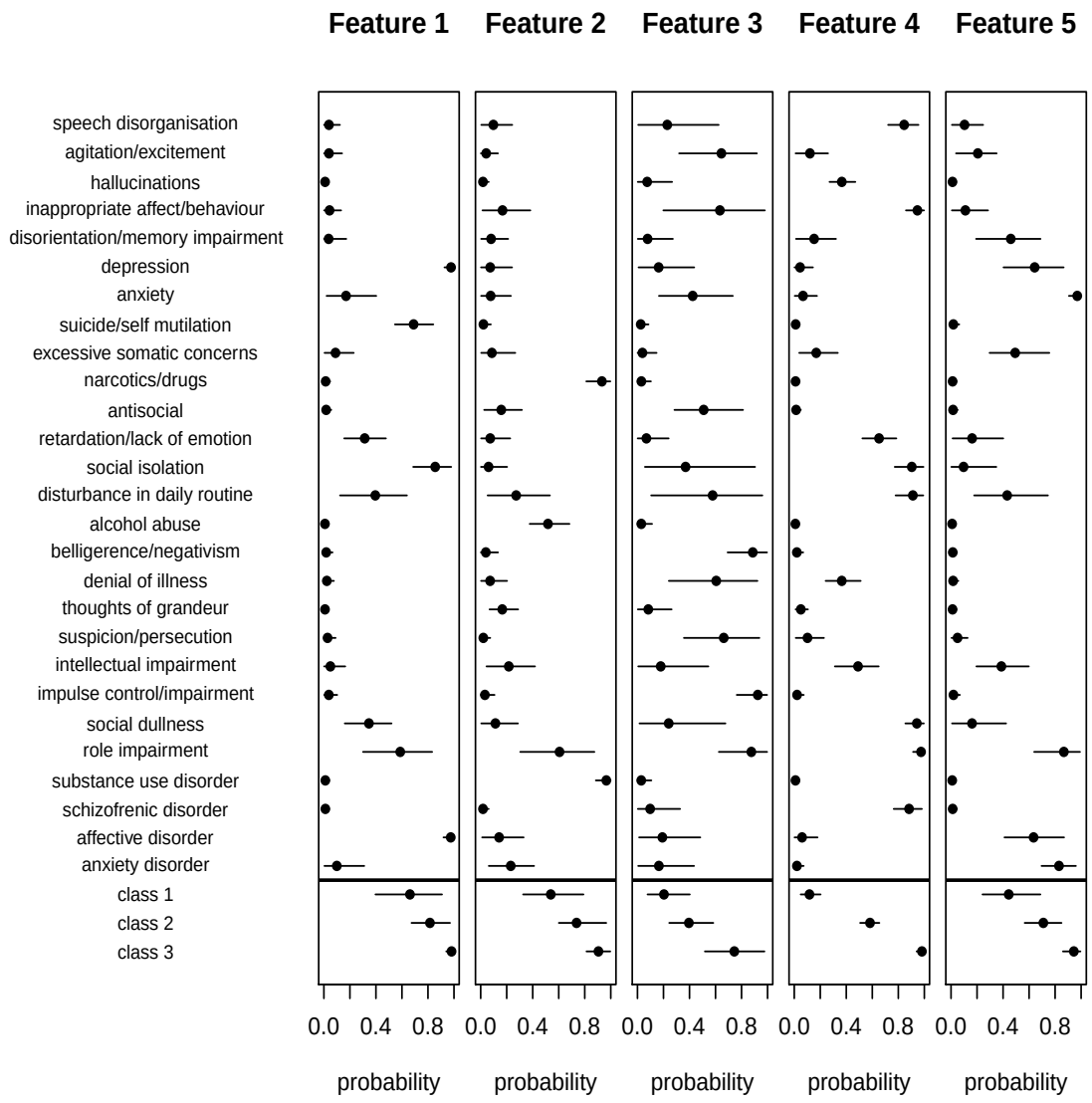

Figure 6. Posterior mean and 95% posterior interval for the symptom parameters $\boldsymbol{\rho}$ (upper part figure) and selection parameters $\boldsymbol{\gamma}$ (bottom part figure) of a three-class five-feature model ($Q = 3, F = 5$) with probabilistic feature selection and varying patient and symptom classification.